Verification of land-atmosphere coupling in forecast models, reanalyses and land surface models using flux site observations

Paul A. Dirmeyer[1]*, Liang Chen[1], Jiexia Wu[1], Chul-Su Shin[1], Bohua Huang[1], Benjamin A. Cash[1], Michael G. Bosilovich[2], Sarith Mahanama[2], Randal D. Koster[2], Joseph A. Santanello[2], Michael B. Ek[3], Gianpaolo Balsamo[4], Emanuel Dutra[5], and D. M. Lawrence[6]

[1]Center for Ocean-Land-Atmosphere Studies, George Mason University

[2]NASA / Goddard Space Flight Center

[3]NOAA / National Centers for Environmental Prediction / Environmental Modeling Center

[4]European Centre for Medium-range Weather Forecasts

[5]Instituto Dom Luiz, Faculdade de Ciências, Universidade de Lisboa[5]National Center for Atmospheric Research

*Corresponding Author:

Paul A. Dirmeyer
Center for Ocean-Land-Atmosphere Studies
George Mason University
4400 University Drive, Mail Stop: 6C5
Fairfax, Virginia 22030 USA
pdirmeye@gmu.edu

Submitted to: *Journal of Hydrometeorology*

Abstract:

We confront four model systems in three configurations (LSM, LSM+GCM, and reanalysis) with global flux tower observations to validate states, surface fluxes, and coupling indices between land and atmosphere. Models clearly under-represent the feedback of surface fluxes on boundary layer properties (the atmospheric leg of land-atmosphere coupling), and may over-represent the connection between soil moisture and surface fluxes (the terrestrial leg). Models generally under-represent spatial and temporal variability relative to observations, which is at least partially an artifact of the differences in spatial scale between model grid boxes and flux tower footprints. All models bias high in near-surface humidity and downward shortwave radiation, struggle to represent precipitation accurately, and show serious problems in reproducing surface albedos. These errors create challenges for models to partition surface energy properly and errors are traceable through the surface energy and water cycles. The spatial distribution of the amplitude and phase of annual cycles (first harmonic) are generally well reproduced, but the biases in means tend to reflect in these amplitudes. Interannual variability is also a challenge for models to reproduce. Our analysis illuminates targets for coupled land-atmosphere model development, as well as the value of long-term globally-distributed observational monitoring.

1. Introduction

Many LSMs were developed and pressed into service during the 1980s and 1990s to provide lower boundary conditions for the atmospheric GCMs used in climate and weather simulation and prediction (Santanello et al. 2017). This occurred at a time when observations of key land surface variables, and the coupled processes that link the water and energy cycles between the land and atmosphere, were extremely limited. As a result, performance of coupled LSM-GCM systems has been sub-optimal (Dirmeyer et al. 2017).

The necessary observational data sets for validation are only recently becoming available; datasets that combine co-located measurements of land surface states, surface fluxes, near-surface meteorology, and properties of the atmospheric column. Early field campaigns (e.g., Sellers et al. 1992, 1995; Famiglietti et al. 1999; Jackson and Hsu 2001; Andreae 2002) provided observations that helped advance theory and model parameterization development, but their short periods of operation meant collected data provided limited sampling of the phase-space of land-atmosphere interactions, rarely quantifying interannual variability. In the mid-1990s, networks of observing stations began to be established and maintained, providing long-term data sets. A growing number of soil moisture monitoring networks have been established. Their data have been collated, homogenized and standardized by two separate efforts (Dorigo et al. 2011, 2013, 2017; Quiring et al. 2016). Those data sets were used by Dirmeyer et al. (2016) in a first-of-its-kind multi-model multi-configuration assessment of soil moisture simulation fidelity.

Simultaneously, efforts began in the ecological community to collect surface flux data over a variety of biomes (FLUXNET; Baldocchi et al 2001). Over time, in consultation with interested scientific communities, FLUXNET expanded their instrumentation suite to measure soil moisture, ground heat flux, and four-component radiation, allowing detailed closure of the surface energy balance. Rigid standards for data formatting and dissemination within and across regional networks was lacking, so a global standardized and quality-controlled subset of data from many FLUXNET sites was produced ("La Thuile FLUXNET dataset", cf. http://www.fluxdata.org) covering multiple links in the coupled land-atmosphere process chain (Santanello et al. 2011). The La Thuile data set enabled a greater degree of model validation (e.g., Williams et al. 2009; Bonan et al. 2012; Boussetta et al. 2013; Melaas et al. 2013; Balzarolo et al. 2014; Purdy et al. 2016).

In this study, we employ the updated FLUXNET2015 synthesis data set, (Pastorello et al. 2017) expanding the multi-model multi-configuration study of soil moisture simulations in Dirmeyer et al. (2016) to a global assessment of surface energy and water balance simulations, and basic metrics of land-atmosphere coupling. Section 2 describes the observational data and models examined. The next three sections present validations of model annual means, annual cycles, and coupling metrics. We then discuss some of the pathological model behaviors that emerge from the analysis and present conclusions. Throughout the paper we present synthesis figures. Detailed scatter plots showing results across all FLUXNET2015 sites for each model are consigned to the Supplement.

2. Data and Models

The range of dates of data varies considerably among model simulations, and also between individual observational sites. We analyze spatial variability and compare only climatologies (annual means or mean annual cycles) in order to minimize the effect of such asynchronicities, and present a quantification of interannual variability. It is not the intent of this study to validate model simulations of specific events, but rather their overall coupled land-atmosphere behavior. Note also that many coupling metrics, including those used here, can be calculated for LSMs from a combination of forcing and model output, even though the LSMs are not coupled to GCMs.

2.1 Observed data

In situ measurements of near surface meteorological variables, surface fluxes and soil moisture used for model validation come from the November 2016 version of the FLUXNET2015 station data set. Daily, monthly and yearly data have been used; processing of the meteorological, radiation, heat flux and surface hydrologic data including gap-filling are described by Reichstein et al. (2005) and Vuichard and Papale (2015). Only the Tier 1 (open access) data are used in this study (see Table S1 for a complete list of sites) – Figure 1 shows the spatial distribution of sites and some of the key characteristics regarding data availability. 166 sites provide 1242 site-years of data, but coverage is concentrated in the mid-latitudes and particular underrepresentation in the tropics.

The variables processed for this analysis include surface pressure, near surface air temperature and vapor pressure deficit, precipitation, four-component and net radiation, surface sensible and latent heat fluxes (gap-filled following the method of Reichstein et al. 2005 and energy balance closure-corrected) and soil water content measured at the first (shallowest) sensor. There is no consolidated information on the depth of the shallowest sensor across all sites, but typically it is at 5cm or 10cm below the surface. Vapor pressure deficit is converted to specific humidity using the Clausius-Clapeyron relationship. We have used the provided FLUXNET2015 data at the corresponding time intervals for each calculation: yearly data for annual means, monthly data for annual cycles, and daily data for calculating coupling indices.

In addition, we examine a number of gridded global precipitation products for comparison to FLUXNET2015 sites. These are listed in Table S2.

2.2 Model systems

Four global modeling systems are evaluated; two from operational forecast centers and two that are primarily used for research. The operational systems are from the U.S. National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Prediction (NCEP) and the European Centre for Medium-range Weather Forecasts (ECMWF). The research systems are from the U.S. National Aeronautics and Space Administration (NASA) Global Modeling and Assimilation Office (GMAO) and the U.S. National Center for Atmospheric Research (NCAR).

Table 1 summarizes the model components and configurations. Generally, each modeling system is interrogated in three different configurations: 1) LSM only (offline), driven by gridded observationally-based meteorological analyses including downward radiation; 2) LSM coupled to GCM in a *free-running* mode where the coupled system evolves unconstrained after initialization; 3) Reanalysis, where the coupled LSM and GCM are constrained by data assimilation at diurnal or sub-diurnal increments to represent the actual historical evolution of state variables. The NCAR model system does not have an associated reanalysis, so to keep the four-by-three matrix filled, two different reanalyses from GMAO are included. Note that when the coordinates for a FLUXNET2015 site lie within a model's ocean grid cell, it is excluded from comparisons for that model. Thus, the number of stations compared vary from model to model depending on resolution and the land-sea mask.

2.2.1 NCEP

Data for the offline configuration comes from an author-produced simulation using Noah LSM version 2.7.1 (Ek et al., 2003, Mitchell, 2005) driven by 3-hourly gridded meteorological data from the Terrestrial Hydrology Research Group at Princeton

University (Sheffield et al., 2006). The free-running coupled land-atmosphere simulation consists of a subset of 48 years from a 420 year long current climate simulation of CFSv2 initialized in 1980 (Shukla et al. 2017). The coupled simulation is unique among the model systems in that it also includes a coupled ocean component. However, this should have very little effect on the local coupled land-atmosphere behavior of the model. Years 2101-2148 of the simulation are used, but the calendar dates have no real meaning in a fully coupled climate model so far from the initial state, wherein attributes such as atmospheric composition, solar intensity, orbital parameters, etc., are held constant at late 20th century values. The latest NCEP reanalysis is also examined (CFSR; Saha et al. 2010), which combines a global land data assimilation system derived from the NASA Land Information System (LIS; Peters-Lidard et al., 2007), driven by a blended global precipitation analysis (Xie and Arkin 1997; Xie et al. 2007), used to update the coupled analysis cycle once per day over the period 1979-2009.

2.2.2 GMAO

Two reanalyses are included for GMAO; version 1 and version 2 of the Modern-Era Retrospective Analysis for Research and Applications (MERRA; Rienecker et al. 2011, Reichle et al. 2017a). MERRA data cover the period 1980-2015. MERRA-2 is the current state-of-the-art reanalysis covering 1980-2015 (Molod et al. 2015, Gelaro et al. 2017), and is the source of most of the meteorological forcing data for the offline simulation of the Catchment LSM v25 C05 (GMAO 2015a,b). As part of the MERRA-2 reanalysis, the GCM-generated precipitation is corrected with observations-based precipitation before it reaches the land surface (Reichle et al. 2017b); the reanalysis meteorological fields thus feel the observed precipitation rates indirectly through the surface fluxes. Additionally, a global 36-year offline Catchment simulation on the MERRA grid and a 16-year coupled

GEOS5-Catchment simulation at half-degree resolution with prescribed observed SSTs were generated for this comparison.

2.2.3 NCAR

There is no operational reanalysis produced with the NCAR Community Earth System Model (CESM). However, CESM is widely used for research in the academic community, and we have generated offline and coupled simulations for this comparison. The offline simulation uses version 4.5 of the Community Land Model (CLM; Lawrence et al. 2011) driven with forcing spanning 1991-2010 from version 4 of the blended and gap-filled CRUNCEP (Viovy 2013) 0.5° data set (available at: https://www.earthsystemgrid.org/dataset/ucar.cgd.ccsm4.CRUNCEP.v4.html) aggregated to the nominal 1° GCM resolution. A simulation with CLM4.5 coupled to CAM4 in CESM1.2.2 has been produced spanning 1991-2014 with specified climatological SSTs.

2.2.4 ECMWF

The offline simulation from ECMWF is with Cycle 43R1 of the Hydrology Tiled ECMWF Scheme of Surface Exchanges over Land (HTESSEL) run at ~16km resolution based on a cubic octahedral global grid (TCo639) for the period 1979-2015. This offline simulation follows ERA-Interim/land configurations closely (see Balsamo et al. 2015), forced by ERA-Interim meteorology and fluxes with an altitude correction applied to temperature, humidity and surface pressure. This offline simulation is used to initialized the land state of the operational ECMWF hindcasts. The coupled simulation comes from the Athena Project (Kinter et al. 2013) for 1961-2007 where an older version of HTESSEL is coupled to IFS Cycle 32R3 at a similarly high native horizontal resolution and specified observed SSTs, but the data has been post-processed to a 1.125° uniform grid. ERA-Interim (Dee et al. 2011), spanning 1979-2015, provides the reanalysis configuration of data for the comparison, which used TESSEL prior to hydrology upgrades.

3. Annual Means

The comparison of models to FLUXNET2015 observations of annual means amounts to an assessment of model ability to reproduce global spatial patterns (within the limitations of the uneven distribution of station locations) of the variables' time averages. For the offline LSM simulations, meteorological forcing data are specified from gridded data sets, so their correlation to FLUXNET2015 observations is not a pure reflection of model performance as the forcing data constrain LSM behavior. Similarly, for the reanalysis products, performance reflects a combination of model characteristics, data assimilation techniques and the distribution and quality of the data assimilated. Assimilation of observational data constrains the coupled land-atmosphere model behavior to some degree. While the free-running model simulations provide an unabridged assessment of model performance, results from the other modes of simulation are nevertheless enlightening.

As an indicator of observational uncertainty and the impact of comparing model grid box values to field sites, we first note how a number of gridded observational precipitation products and the reanalyses validate against precipitation measurements at FLUXNET2015 locations. Figure 2 shows mean (dots) and span (whiskers) of annual precipitation totals, where the abscissa always corresponds to measurements from the FLUXNET2015 sites. For most sites, the observational products (top two rows of Fig. 2) cover the entire time span of FLUXNET2015 observations (see Table S2 for details). All reanalyses (bottom row of Fig. 2) except CFSR span the FLUXNET2015 period. Several statistics of spatial agreement are shown: Pearson's product moment correlation coefficient (r_p), Spearman's rank correlation coefficient (r_s), root mean square error (RMSE), slope of the best-fit linear regression of Y on X (Slope) and the fraction of total stations (labeled "Span Diag" in Fig. 2) where the span of the individual annual totals from the gridded products (vertical whiskers) overlap the span from FLUXNET2015 sites (horizontal whiskers). The last statistic tests the possibility that the FLUXNET2015 observations and gridded estimates do not come from distinct populations, i.e. their ranges overlap.

Estimates from gridded observational data sets, which range in spatial resolution from 0.25° (MSWEP, TRMM) to 2.5° (GPCP), provide a plausible upper bound to the accuracy we could expect from gridded Earth system models. For the 166 (or fewer) FLUXNET2015 sites compared, which admittedly represent a rather uneven sampling of global terrestrial precipitation, three observational products score at the top: MSWEP, CPC-Uni and U.Del. Each has a Pearson's correlation of nearly 0.8, a rank correlation between 0.8-0.9, and the highest number of stations whose ranges span the diagonal X=Y line. The lower limit for RMSE across these sites is about 240mm. Note that all gridded products underestimate the slope, indicating the inability of large area averages to resolve local variations in average precipitation.

MERRA-2 performs on par with the best gridded observed products, namely because it reports a bias corrected precipitation that is used as part of the assimilation process instead of model-generated precipitation as an input to the LSM (Reichle and Liu 2014). Thus, it is effectively another gridded observational data set for precipitation. Figure S1 compares the precipitation predicted by the model physical parameterizations in MERRA-2 alongside the corrected version in the same fashion as Fig 2. The correction greatly reduces bias, cuts RMSE by one third, slightly improves spatial correlations, and increases the number of stations spanning the diagonal by 28%. CFSR significantly underperforms other reanalyses at FLUXNET2015 locations.

Precipitation is among the most difficult quantities for models to simulate. We expect among near surface meteorological variables the lowest correlations and largest coefficient of variation for precipitation. It also has many observationally-based data sets to choose from, providing a robust estimate of skill to be expected from comparing point measurements to gridded data sets. Figure 2 provides generous thresholds, particularly for correlations, to keep in mind when assessing model simulations of the terms of the surface water and energy balance. As shown below, correlations of 0.7-0.8 are a challenge for models to attain for precipitation, as well as some other water and energy budget terms.

Among near surface meteorology (e.g., temperature and specific humidity) and downward surface fluxes (including shortwave and longwave radiation), precipitation has the greatest small-scale variability on monthly to annual time scales, and is thus the most difficult land surface "forcing" to replicate at the FLUXNET2015 sites. Figures S2-S6 show the scatters and statistics for the models listed in Table 1 for these five variables. Here, the restriction that the years of the models match those at each FLUXNET2015 site is lifted, and the climatologies of the complete data sets are compared. Not surprisingly, the global distribution of annual mean temperature is very well reproduced by the models (Fig. S2), with 88-96% of the observed variance explained. Observed specific humidity is only slightly less well correlated among the models (Fig. S3), but there is a consistent positive bias relative to FLUXNET2015 measurements. Patterns of annual mean downward radiation (Figs. S4 and S5) are well simulated, with a tendency for a slight negative bias in longwave radiation (Fig. S5), and a stronger positive bias in shortwave radiation across models (Fig. S4), consistent with other assessments of model shortwave errors that depend on GCM radiative transfer parameterizations (cf. Slater 2016). Precipitation shows the least agreement; note the bottom row of Fig. S6 is not identical to that of Fig. 2 because the years compared differ. Nevertheless, the results are similar. We can consider MERRA-2 as representing the upper limit of comparison for annual precipitation when the periods do not match between models and observations. Offline Catchment actually performs slightly better than MERRA-2, and CFSv2 is generally the poorest performing model system in the set. Free-running climate models understandably perform worse than either reanalyses or offline LSM simulations, as they are least constrained by observational data. In the case of CFSv2, there are essentially no constraints within the Earth system as an ocean model is coupled; other free-running simulations have specified SSTs.

Precipitation is a major source of error at the land surface, but so are elements of the radiation budget. We employ Taylor diagrams to synthesize the statistics of correlation across FLUXNET2015 sites; RMSE and standard deviation are normalized by observed values. Figure 3 shows the global distribution of annual mean downward radiation terms is well simulated across all model configurations, with downward shortwave radiation performing slightly better than downward longwave radiation. Recall for the LSM-only models, downward radiation is an input forcing, and the quality of those data sets can vary significantly (Slater 2016). However, the distribution of upward shortwave radiation is rather poorly simulated, with the NCEP models showing the worst correlations, and the NCAR models the best (yet explaining less than half of the variance). There is also a strong tendency to under-represent the spatial variability (normalized standard deviations less than 1) of downward shortwave radiation. This degrades simulation of net radiation, which has consistently lower correlations than downward radiation terms, yet uniformly better than upward shortwave radiation. The overlap of the spans of annual mean values from models and observations (size of the dots) generally decrease from shortwave down to longwave down to shortwave up.

Figure 3 implies discrepancies in the representation of surface albedo across models at FLUXNET2015 sites. We show a Taylor diagram for calculated albedo in Fig. 4. As there are many sites at relatively high northern latitudes that experience snow cover for some part of the year, snow albedo could specifically be a problem. However, a plot of only the JJA albedo verification shows boreal summer generally has even lower fidelity, and systematically low spatial variability, compared to the annual mean. The overlap between the spans of annual mean albedos range among the models from 16% to 38% of FLUXNET2015 sites, but for JJA they span only 13-24%.

The low variability could be explained by the fact that most LSMs, whether stand-alone or coupled, have a simple parameterization of albedo based on properties of a small number of vegetation and soil types, often specified as a climatological seasonal cycle. CLM actually calculates surface albedo based on a number of properties including vegetation density and zenith angle of the sun, which may lead to the somewhat better performance of the NCAR models. As described later, the offline NCEP LSM (identified as NL) specifies a multi-year satellite-derived monthly green vegetation fraction as a boundary condition that appears in Fig. 4 to enhance variability, while its positive biases have been noted by Xia et al. (2012). Furthermore, discrepancies between grid box average albedo and local conditions at field sites, including the effect of vegetation differences and soil moisture on albedo (Zaitchik et al. 2013), could add spatial "noise" to the FLUXNET2015 values relative to what models are representing. Nevertheless, such discrepancies lead to a degradation in the representation of surface available energy that is partitioned between sensible, latent and ground heat fluxes. Even an otherwise "perfect" LSM could not produce the right values of these fluxes if net radiation is incorrect. Coupled with errors in precipitation, which affect available soil moisture and thus Bowen ratios, LSMs are at a compounded disadvantage in simulating the surface water and energy budget terms.

In Fig. 5 we correlate across the stations the mean errors in key water and energy cycle quantities and present a schematic representation of the relative coupling or connectedness exhibited between terms. This also suggests how errors in the simulation or specification of one term can propagate to others through the land-atmosphere coupling process chain (cf. Santanello et al. 2011). r_s is generally larger than r_p because it does not overemphasize outliers, thus is used for this comparison. Ratios show the fraction of models with correlations at the 90% confidence level, and p-values are based on the average correlation across models. Note the number of included stations varies depending on the availability of observed data (recall from Fig. 1 that a number of FLUXNET2015 sites do not allow for albedo estimations) and among models depending on whether the corresponding grid box is water or land. Furthermore, the data saved from the free-running ECMWF model simulations (EC) do not allow for estimation of albedo, so 11 models are compared for albedo.

Unsurprisingly, we find surface net radiation errors correlate strongly to albedo errors, with 11 of 11 models registering significant correlations (two-tailed p-values < 0.05) and the multi-model average correlation across 114-118 sites has a p-value of $4\text{x}10^{-7}$. For net radiation versus precipitation, only 2 of 12 models (CL and M1) show significant correlation across 144-151 sites and p=0.55 for the multi-model average, so no direct arrow is drawn in Fig. 5. Note that precipitation errors arise not only from misrepresentation of land-atmosphere interactions, but also from the parameterization of dynamic and thermodynamic processes (so-called “model physics”) in the GCM.

FLUXNET2015 reports both raw and Bowen-ratio corrected heat fluxes. Corrected fluxes are available at fewer than 100 of the sites (two-tailed p=0.05 for correlations $|r| \gtrsim 0.2$, compared to $|r| \gtrsim 0.16$ for the full set of sites), but generally correspond better to the models than uncorrected fluxes, which do not close the surface energy balance (cf. Figs. S9-S12). Regardless, the same story emerges with either set of fluxes: precipitation errors correlate significantly to latent heat flux errors (p=0.02 in Fig. 5) but not sensible heat flux errors (p=0.31). Meanwhile, albedo errors are very strongly linked to sensible heat flux errors ($p=7 \times 10^{-5}$) but not latent heat flux errors (p=0.69). Evaporative fraction (EF; the fraction of sensible + latent heat flux accounted for by the latent heat flux) relates strongly to both, but more strongly to errors in albedo (p=0.003) than precipitation (p=0.05). Consistently, correlating EF errors to the heat flux errors (black two-way arrows) demonstrates more variance explained by sensible heat flux than latent heat flux. Finally, LCL errors relate strongly to precipitation errors ($p=2 \times 10^{-5}$) but are marginally significant in relation to albedo errors (p=0.06). LCL has a prevalent negative bias (Fig. S8) reflecting the positive biases in specific humidity.

This analysis shows that models have troublesome errors in both the surface water and energy cycles, which make their way into the land-atmosphere coupling process chain. As a result, the degree to which weather and climate models correctly simulate feedbacks of land surface anomalies onto the atmosphere may be cast into some doubt. However, the origins of several sources of error have been identified and their alleviation can be pursued. In section 5 we will examine directly model fidelity in simulating metrics of land-atmosphere coupling.

4. Mean Annual Cycle

The next criterion for models, beyond simulating the annual means among FLUXNET2015 sites, is reproducing the annual cycle. The first harmonic is fit to the 12 monthly means for each variable, determining phase and magnitude (half of valley-to-peak distance) using a standard Fourier transform. Errors in phase and magnitude at each station, quantified across all stations with similar metrics as the annual mean, indicate skill in simulating the annual cycle. Amplitude errors are displayed in conventional scatter diagrams (see Figs. S15-S24), but to display information for phase errors, we have configured the classical scatter diagram in a polar projection (see Figs. S25-34; the caption of Fig. S25 gives a detailed description of those plots). The whiskers in the supplemental figures again show models frequently display a smaller range of year-to-year variability than data from FLUXNET2015 sites. This may be partially explained by the scale difference (point measurements will vary more than grid-box averages) but is also likely due to the overly deterministic nature of many model parameterizations (Palmer 2012).

Taylor diagrams summarize the results across models. We focus on depictions of energy budget terms, as they reveal some of the main issues among models. Figure 6 shows model performance in simulating the amplitudes of the annual cycles of net radiation, sensible and latent heat fluxes across FLUXNET2015 sites. All model products demonstrate similar skill for net radiation, clustered between 0.64-0.78 correlation and a tendency toward too large an annual cycle. Only the offline NCEP and coupled ECMWF models have a negative bias in amplitude. Latent heat flux simulations show lower skill for every model, clustering between 0.28-0.43 for correlations. At the stations where energy balance corrected fluxes are provided, correlations improve to 0.37-0.50 (not shown). The positive bias is not so pervasive for latent heat; rather it appears the positive bias in net radiation tends to be expressed in the sensible heat term. There is also a much larger spread among models for sensible heat, both in terms of correlation (0.14-0.54) and normalized standard deviation (0.78-1.50).

The models' skill in representing the phase of the annual cycle has a similar distribution (Fig. 7). The phase of net radiation is best represented, latent and sensible heat have spatial correlations of phasing between ~0.8-0.92 with sensible heat phases having slightly lower fidelity in general. It is interesting as the general consensus is that sensible heat flux is a simpler process to model than latent heat flux, yet it has been shown in other contexts that LSMs struggle more to simulate sensible heat flux (e.g., Best et al. 2015).

The Taylor diagram for the annual cycle of albedo (Fig. 8) shows very similar correlations of the yearly amplitude between models and observations (0.50-0.71) but a large range in standard deviation; Noah v2.7.1 (NL) shows a particularly high value contributing to large RMSE. The phase is better represented by all models, but interestingly the standard deviations are uniformly over-estimated. Most models now use global MODIS-based data sets of albedo as either a parameter set or for calibration of surface radiative parameterizations, so the large inter-model spread and lack of obvious clustering within families of models is surprising.

5. Coupling Metrics

Correlations between land surface state variables and surface fluxes (the terrestrial leg of coupling) and between land surface fluxes and atmospheric states or properties (atmospheric leg) may indicate feedbacks. For instance in the terrestrial leg, positive (negative) correlation between soil moisture and latent (sensible) heat flux implies soil moisture control of fluxes (a moisture limited situation) as opposed to energy (net radiation) limited situations where atmospheric states control the fluxes. However, the variance in the driving term(s) must also be sufficiently large for a sensitivity of atmosphere to the land to have a consequential impact on climate, relative to other factors. A coupling index I can be constructed from terms in either leg: $I = \sigma(b)r(a,b) = \sigma(a)\frac{db}{da}$ where a is the forcing and b is the responding variable, σ is standard deviation in time, r is correlation in time, and the linear regression slope of b on a is a measure of the sensitivity of b to a (Dirmeyer 2011, Dirmeyer et al. 2013).

Figure 9 synthesizes the performance of the various model configurations regarding two-legged coupling metrics linking soil moisture to boundary layer properties. The formulae for the coupling indices are indicated on the figure axes calculated from daily mean values. The terrestrial leg quantifies the combined sensitivity (correlation) of surface fluxes (here, latent heat flux) to land states (soil moisture) with variability (standard deviation) of the flux. The atmospheric leg links surface fluxes (sensible heat flux) to atmospheric states (LCL, which combines near surface temperature and humidity information). Larger values denote stronger feedback linkages.

In each panel of Fig. 9, similar to the approach of Sippel et al. (2017), quantities are calculated for the three consecutive months that have the warmest average temperature according to the FLUXNET2015 data. We distinguish between positive values of each metric, which indicate the existence of feedbacks from land to atmosphere, from negative (no feedbacks) by coloring the four quadrants by their coupling regimes: red = both legs present and a full coupling pathway; green = the land leg is present, the atmospheric leg is missing; blue = atmospheric leg is present, land is missing; grey = neither leg present. The white dots show where FLUXNET2015 sites fall in this two-dimensional metric space. The colored dots are each model's rendering of the metrics for the grid boxes containing the FLUXNET2015 sites; the color indicates the quadrant according to the FLUXNET measurements. Thus, the more colored dots that fall in the quadrant with the matching color, the better the model is reproducing the global pattern of coupling regimes.

The model centroid usually lies below and to the right of the observed centroid for a given coupling regime, meaning models tend to over-estimate the terrestrial coupling index (the rightward offset), yet underestimate the strength of the atmospheric leg (the downward offset). Recall the number of FLUXNET2015 sites compared is not the same for each model. The percentage in each quadrant indicates how many of the FLUXNET2015 sites in that regime are correctly placed in the right quadrant. For instance, the CFS Reanalysis has 76% of the FLUXNET stations exhibiting both coupling legs (red) in the correct regime. However, there are clearly many dots of other colors also in the red quadrant, showing the model places many other stations erroneously in that regime. Interestingly, none of the models put the few sites with no warm-season coupling in the grey quadrant. Overall, the reanalyses perform best: a 56.5% overall hit rate for the fully-coupled regime versus 52.8 for coupled models, and 44.0% for offline LSMs; and for the atmosphere-only coupling regime 49.2% versus 33.0% for coupled models and 31.6% for offline LSMs.

We have also examined performance of the models for their simulation of the observed FLUXNET2015 correlations and standard deviations (the two terms in the coupling indices) separately. As implied previously for the terrestrial leg, there is a positive bias in correlations for all models except for ERA-Interim (Table 2). Bias in the standard deviation of latent heat fluxes across all sites is small for most models, so most of the positive bias in coupling index comes from the correlation term. The model biases are even stronger in the anti-correlation between soil moisture and sensible heat flux (not shown). However, there is generally an even greater bias in correlations for the atmospheric leg (Table 2) paired in every model with an underrepresentation of the daily variability of the LCL. These two biases compound, leading to the strong underrepresentation of coupling in the atmospheric leg of land-atmosphere interactions.

There are several caveats to note. First, the notion of calculating the atmospheric coupling leg from offline LSM simulations is only partially justifiable. It is certainly possible to calculate the correlations between surface fluxes and LCL height (which depends on near-surface meteorological data supplied as forcing to the LSM), but there is no possibility for the fluxes to affect 2m temperature or humidity. Thus, this is more of a test of model consistency than a true diagnosis of coupling.

Second, estimates of the correlation component of the coupling indices from observed data must be closer to zero than the true values in nature, because random measurement errors will degrade correlations (Robock et al. 1995). Thus, it is not necessarily wrong that models show a stronger terrestrial coupling leg than FLUXNET2015 data. The degree of impact can be estimated for variables such as soil moisture, whose auto-correlation time scales are much longer than the daily data interval (cf. Dirmeyer et al. 2016) but can be difficult to estimate from small samples or for other quantities. Nevertheless, the fact that models routinely underestimate the strength of the atmospheric leg runs counter to being attributable to random observational errors at FLUXNET sites, and likely represents real model bias.

Finally, the difference in scale between flux tower measurements (typically representative of conditions in an area of a square kilometer or less) and model grid-box averages (here ranging from 200–2x10^4 km^{-2}) can affect statistics. Dirmeyer et al. (2016) showed there was little sensitivity of estimates of temporal variations in daily soil moisture to spatial scale differences in the model grid box range, however, the same may not be true for other terms, or for correlations. The larger the averaging area, the smoother we should expect time series to be, potentially affecting estimation of coupling indices.

6. Discussion and Summary

We have confronted four different global model systems in multiple configurations (LSM only, LSM coupled to GCM, and reanalysis) with flux tower observations from 166 sites in the global FLUXNET2015 data set to determine how well they reproduce the spatial distribution of annual means and the annual cycle of state variables and terrestrial surface fluxes, and coupling indices between land and atmosphere. Returning to Table 2, there is a separation evident between the three classes of models. For the terrestrial leg of land-atmosphere coupling, all models appear to overestimate correlations between soil moisture and latent heat flux, with the caveat discussed previously that correlations necessarily skew low when calculated from observed data. Nevertheless, assuming as much as a 50% reduction from true correlations, it appears the reanalyses do the best job at reproducing observed correlations, followed by the free-running models and last the uncoupled LSMs. There is a similar stratification for the standard deviation of latent heat flux: reanalyses very closely represent the observed temporal variability of this flux, while coupled models and stand-alone LSMs progressively underestimate it. For the atmospheric leg, represented by the coupling index between sensible heat flux and LCL height, all classes of models severely underestimate the correlation and the day-to-day variability in the LCL. Reanalyses again do the best job at correlations and stand-alone LSMs are the worst. Here, coupled models fare slightly better than reanalyses in representing LCL variance. Given that reanalyses are somewhat constrained by the assimilation of observations, the errors in those models do not manifest as freely, so it makes sense reanalyses should verify the best. On the other hand, offline LSMs lack some of the coupling we are trying to gauge. For example, surface sensible and latent heat fluxes cannot affect near surface temperature and humidity in such a configuration. This prescription of near-surface states interferes with the feedback processes.

General characteristics of note are that scatter diagrams of model versus FLUXNET2015 quantities almost always show a linear regression slope indicating a wider range of variation in the observations. Models also tend to have lower interannual variability (length of whiskers) than observations suggest. These traits are consistent with scale differences between model grid cells and the area sampled by flux towers; model grid values represent areas at least 2-4 orders of magnitude larger, which particularly affects precipitation forcing. Thus, this difference is not a concern regarding model performance *per se*, but rather representativeness across scales.

Another general characteristic is that the models verify better against the corrected surface fluxes and quantities derived from them; wherein observed sensible and latent heat values are adjusted to close the surface energy budget. This makes sense as models close surface energy (and water) budgets by design, whereas closure is not assured in an observational setting where a number of instruments, with different calibrations and error characteristics, contribute separate terms of the surface balances. However, when the propagation of model errors through the energy and water cycles are traced (Fig. 5), EF in models shows strong sensitivity to radiation errors, implying that conservation of Bowen ratio (and thus EF) as a means to correct observed heat fluxes and close the energy balance may not be the most efficacious.

There are differences that do appear to reflect general model biases. All models and configurations show a positive bias in near-surface humidity (Fig. S3, S14), downward shortwave radiation (Figs. S4, S17) and a range of biases in downward longwave radiation (Fig. S5). Such radiation biases are a long-standing problem in global models (cf. Dirmeyer et al. 2006), and stem from problems in the parameterization of atmospheric radiative transfer, clouds and aerosols in GCMs. However, not all radiative errors are atmospheric in origin – there is clear indication that LSMs struggle to represent the spatial and temporal variability of surface albedo (Figs. 4, 8).

Combined with well-known difficulties models have in simulating precipitation (Figs. 2, S6, S15, S25), it becomes extremely challenging for models to partition available energy correctly at the surface between latent, sensible and ground heat fluxes, and to reproduce the spatiotemporal patterns of relationships between soil moisture, surface fluxes and the lower troposphere. Errors in latent heat flux generally correlate significantly to precipitation errors, while sensible heat flux errors relate strongly to surface albedo errors. Evaporative fraction errors connect to both, but more strongly to the energy (albedo – sensible heat flux) pathway than the water (precipitation – latent heat flux) pathway. Height of the LCL, which has a strong negative bias across all models related to the positive humidity bias, has errors that correlate strongly to the water cycle pathway, but also to the energy cycle pathway.

The spatial distributions of the annual cycles are generally well reproduced for energy budget terms, except for upward shortwave radiation, related to the albedo problems discussed earlier. However, there is a tendency for too strong a seasonal cycle in net radiation, caused by excessive summertime downward shortwave radiation, and expressed more strongly in the annual cycle of sensible heat flux than latent heat flux. Models generally do very well representing the spatial distribution of the phasing of the annual cycle, even for precipitation (64-92% of variance explained) and soil moisture (40-61% of variance explained).

Finally, despite the barriers described above to models' capacity to represent the spatiotemporal distribution of land-atmosphere coupling, we find models often do a reasonable job. Some systematic biases are evident: models consistently over-estimate the strength of the terrestrial leg of coupling (namely, too strong a correlation between soil moisture and sensible heat fluxes), yet even more clearly underestimate the strength of the atmospheric leg (both the correlation between surface fluxes and boundary layer properties, and day-to-day variability of boundary layer properties). Random observational error tends to reduce correlations between observed quantities, so it is possible that models are not greatly overestimating the terrestrial leg of coupling, or perhaps are not overestimating it at all. However, we find the time series at most FLUXNET2015 sites are too short to robustly estimate the random error effects on correlation – perhaps in another ten years we will be able to quantify these errors. Similarly, the spatial scale differences between observations and model output may contribute to the variance differences in the atmospheric leg, but disparity in correlations between surface fluxes and LCL could only be stronger than calculated here, not weaker, because of the effect of measurement error.

LSMs forced by global gridded meteorology rather than local forcing from the tower sites themselves are handicapped to some degree (cf. Chen et al. 2017). So our most confident conclusion regarding land-atmosphere coupling is that models under-represent the feedback of surface fluxes on boundary layer properties at FLUXNET2015 sites. We find this unique data set has potential for model development and parameter optimization to alleviate biases in model configurations shown to mirror those used in forecasting applications (Orth et al. 2016, 2017).

Overall, we conclude that many of the long-known problems and biases in global models of the land-atmosphere portion of the climate system still exist. Nevertheless, there is a fair degree of compensation among errors, such that model representations of land-atmosphere coupling often appear fairly good. Some targets for model improvement are clear, however, as coupling linkages suggest processes where problems may lie. The representation of surface albedo (LSM) and the quantities of downward radiation at the surface (GCM) need improvement among the energy cycle terms, along with the partitioning of available energy between latent and sensible heat flux (a coupled model development problem). Precipitation errors remain large, and inconsistencies in representing soil moisture among models and between models and nature (cf. Koster et al. 2009) remain stubborn issues.

As one might expect, reanalyses tend to perform better, as they are more constrained by observational data. LSMs run offline also benefit from meteorological forcing that is highly observational in origin, but can be handicapped by their lack of two-way interaction with the lower troposphere. It should be clear from the various figures that individual models perform better or worse at simulating specific facets of land-atmosphere interactions. However, we emphasize here the commonalities among models more than differences. This study is not primarily intended as a model inter-comparison, but rather a multi-model attempt to draw model-independent conclusions about the current state of performance of land-atmosphere models (in various configurations) by confronting them with a new and unique observational data set.

Furthermore, this study is not a final judgement, but a first look that will hopefully catalyze accelerated development and improvement in coupled land-atmosphere modeling. Application of cross-component metrics like coupling indices can reveal prime areas for model development that are not evident from piecewise evaluation of model components. The next step is intensive, focused sensitivity studies with individual models, preferably validated in the context of coupled model systems, that will zero in on the problematic parameterizations. We may also need to revisit some of the fundamental assumptions that underpin the formulations in models (e.g., Cheng et al. 2017).

Furthermore, it is clear that long-term observational monitoring is highly valuable, and that value only increases with the duration of data sets at individual sites. Greater spatial distribution of flux tower sites, especially into under-monitored regions outside middle- and high-latitudes, would further increase the overall usefulness to model development.

Acknowledgments: This work has been primarily supported by National Aeronautics and Space Administration grant NNX13AQ21G. NCAR model simulations were conducted with support from the National Science Foundation grant AGS-1419445. The ERA-Interim reanalysis data are provided by ECMWF and processed by LSCE. This work uses eddy covariance data acquired and shared by the FLUXNET community (listed in Table S1), including these networks: AmeriFlux, AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada, GreenGrass, ICOS, KoFlux, LBA, NECC, OzFlux-TERN, TCOS-Siberia, and USCCC. The FLUXNET eddy covariance data processing and harmonization was carried out by the European Fluxes Database Cluster, AmeriFlux Management Project, and Fluxdata project of FLUXNET, with the support of CDIAC and ICOS Ecosystem Thematic Center, and the OzFlux, ChinaFlux and AsiaFlux offices. Taylor diagrams were produced using a modified version of the GrADS script developed by Bin Guan. We thank Cristina Benzo for her contributions to produce Table S1, and Eleanor Blyth and two anonymous reviewers for their helpful review comments.

References:

Andreae, M. O., and co-authors, 2002: Biogeochemical cycling of carbon, water, energy, trace gases, and aerosols in Amazonia: The LBA-EUSTACH experiments. *J. Geophys. Res.*, **107**, LBA33-1–LBA33-25.

Baldocchi, D., and co-authors, 2001: FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor and energy flux densities. *Bull. Amer. Meteor. Soc.*, **82**, 2415-2434.

Balsamo, G., and co-authors, 2015: ERA--Interim/Land: a global land surface reanalysis data set, *Hydrol. Earth Syst. Sci.*, **19**, 389- 407, doi: 10.5194/hess-19-389- 2015.

Balzarolo, M., and co-authors, 2014: Evaluating the potential of large-scale simulations to predict carbon fluxes of terrestrial ecosystems over a European Eddy Covariance network, *Biogeosci.*, **11**, 2661-2678, doi:10.5194/bg-11-2661-2014.Best, M. J., and Co-authors, 2015: The plumbing of land surface models: benchmarking model performance. *J. Hydrometeor.*, **16**, 1425-1442, doi: 10.1175/JHM-D-14-0158.1.

Bonan, G. B., K. W. Oleson, R. A. Fisher, G. Lasslop, and M. Reichstein, 2012: Reconciling leaf physiological traits and canopy flux data: Use of the TRY and FLUXNET databases in the Community Land Model version 4, *J. Geophys. Res.*, **117**, G02026, doi: 10.1029/2011JG001913.

Boussetta, S., G. Balsamo, A. Beljaars, T. Kral, L. Jarlan, 2013: Impact of a satellite-derived leaf area index monthly climatology in a global numerical weather prediction model. *Int. J. Remote Sens.*, **34**, 3520-3542. doi: 10.1080/01431161.2012.716543.

Chen, L., P. A. Dirmeyer, Z. Guo and N. M. Schultz, 2017: Pairing FLUXNET sites to validate model representations of land use/land cover change. *Hydrol. Earth Sys. Sci. Discus.*, doi: 10.5194/hess-2017-190.

Cheng, Y., C. Sayde, Q. Li, J. Basara, J. Selker, E. Tanner, and P. Gentine, 2017: Failure of Taylor's hypothesis in the atmospheric surface layer and its correction for eddy-covariance measurements. *Geophys. Res. Lett.*, **44**, 4287–4295, doi: 10.1002/2017GL073499.

Dee, D. P., and co-authors, 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553-597, doi: 10.1002/qj.828.

Dirmeyer, P. A., R. D. Koster, and Z. Guo, 2006: Do global models properly represent the feedback between land and atmosphere? *J. Hydrometeor.*, **7**, 1177-1198, doi: 10.1175/JHM532.1.

Dirmeyer, P. A., 2011: The terrestrial segment of soil moisture-climate coupling. Geophys. Res. Lett., 38, L16702, doi: 10.1029/2011GL048268.

Dirmeyer, P. A., S. Kumar, M. J. Fennessy, E. L. Altshuler, T. DelSole, Z. Guo, B. Cash and D. Straus, 2013: Model estimates of land-driven predictability in a changing climate from CCSM4. J. Climate, 26, 8495-8512, doi: 10.1175/JCLI-D-13-00029.1.

Dirmeyer, P. A., and co-authors, 2016: Confronting weather and climate models with observational data from soil moisture networks over the United States. *J. Hydrometeor.*, **17**, 1049-1067, doi: 10.1175/JHM-D-15-0196.1.

Dirmeyer, P. A., P. Gentine, M. B. Ek, and G. Balsamo, 2017: Land Surface Processes Relevant to S2S Prediction. [Chapter 8 in: *The Gap Between Weather and Climate Forecasting: Sub-Seasonal to Seasonal Prediction* (A. W. Robertson and F. Vitart Eds.)], Elsevier, (in revision).

Dorigo, W. A., and co-authors, 2011: The International Soil Moisture Network: a data hosting facility for global in situ soil moisture measurements, *Hydrol. Earth Syst. Sci.,* **15**, 1675-1698, doi: 10.5194/hess-15-1675-2011.

Dorigo, W.A., and co-authors, 2013: Global automated quality control of in situ soil moisture data from the International Soil Moisture Network. *Vadose Zone J.*, **12**(3), doi: 10.2136/vzj2012.0097.

Dorigo, W., and co-authors, 2017: ESA CCI Soil Moisture for improved Earth system understanding: state-of-the art and future directions, *Remote Sens. Env.* (in press), 10.1016/j.rse.2017.07.001.

Ek, M. B., K. E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J. D. Tarplay, 2003: Imple- mentation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *J. Geophys. Res.*, **108**, 8851, doi: 10.1029/2002JD003296.

Famiglietti, J. S., J. A. Devereaux, C. A. Laymon, T. Tsegaye, P. R. Houser, T. J. Jackson, S. T. Graham, M. Rodell, and P. J. van Oevelen, 1999: Ground-based investigation of soil moisture variability within remote sensing footprints during the Southern Great Plains 97 (SGP97) hydrology experiment. *Water Resour. Res.*, **35**, 1839-1851.

Gelaro, R., and co-authors, 2017: The Modern-Era Retrospective analysis for Research and Applications, version 2 (MERRA-2). *J. Climate*, **30**, 5419-5454, doi: 10.1175/JCLI-D-16-0758.1.

Global Modeling and Assimilation Office (GMAO), 2015: MERRA-2 inst1_2d_lfo_Nx: 2d, 1-Hourly, Instantaneous, Single-Level, Assimilation, Land Surface Forcings V5.12.4, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed 3 July 2016, doi: 10.5067/RCMZA6TL70BG.

Global Modeling and Assimilation Office (GMAO), 2015: MERRA-2 tavg1_2d_lfo_Nx: 2d, 1-Hourly, Time-Averaged, Single-Level, Assimilation ,Land Surface Forcings V5.12.4, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed 3 July 2016, doi: 10.5067/L0T5GEG1NYFA.

Jackson, T. J., and A. Y. Hsu, 2001: Soil moisture and TRMM microwave imager relationships in the Southern Great Plains 1999 (SGP99) experiment, IEEE Trans. *Geosci. Remote Sens.*, **39**, 1632-1642.

Kinter III, J. L., and co-authors, 2013: Revolutionizing climate modeling – Project Athena: A multi-institutional, international collaboration. *Bull. Amer. Meteor. Soc.*, **94**, 231–245, doi: 10.1175/BAMS-D-11-00043.1.

Koster, R. D., Z. Guo, P. A. Dirmeyer, R. Yang, K. Mitchell, and M. J. Puma, 2009: On the nature of soil moisture in land surface models. *J. Climate*, **22**, 4322–4335, doi: 10.1175/2009JCLI2832.1.

Lawrence, D. M., and co-authors, 2011: Parameterization improvements and functional and structural advances in version 4 of the Community Land Model. *J. Adv. Model. Earth Syst.*, **3**, doi: 10.1029/2011MS000045.

Mahanama, S. P. P., R. D. Koster, G. K. Walker, L. L. Takacs, R. H. Reichle, G. De Lannoy, Q. Liu, B. Zhao, and M. J. Suarez, 2015: Land Boundary Conditions for the Goddard Earth Observing System Model Version 5 (GEOS-5) Climate Modeling System - Recent Updates and Data File Descriptions. NASA/TM–2015-104606, Vol. 39, 55 pp. Document (4608 kB).

Melaas, E. K., A. D. Richardson, M. A. Friedl, D. Dragoni, C. M. Gough, M. Herbst, L. Montagnani, and E. Moors, 2013: Using FLUXNET data to improve models of springtime vegetation activity onset in forest ecosystems. *Ag. Forest Meteor.*, **171-172**, 46-56.

Mitchell, K., 2005: The Community Noah Land Surface Model User's Guide Public Release Version 2.7.1, [available at: http://www.ral.ucar.edu/research/land/technology/lsm/noah/Noah_LSM_USERGUIDE_2.7.1.pdf].

Molod, A., Takacs, L., Suarez, M., and Bacmeister, J., 2015: Development of the GEOS-5 atmospheric general circulation model: evolution from MERRA to MERRA2, *Geosci. Model Dev.*, **8**, 1339-1356, doi: 10.5194/gmd-8-1339-2015.

Orth, R., E. Dutra, and F. Pappenberger, 2016: Improving weather predictability by including land surface model parameter uncertainty. *Mon. Wea. Rev.,* **144**, 1551–1569, doi: 10.1175/MWR-D-15-0283.1.

Orth, R., Dutra, E., Trigo, I. F., and Balsamo, G., 2017: Advancing land surface model development with satellite-based Earth observations, *Hydrol. Earth Syst. Sci.*, **21**, 2483-2495, doi:10.5194/hess-21-2483-2017.

Palmer, T. N., 2012: Towards the probabilistic Earth-system simulator: a vision for the future of climate and weather prediction. *Quart. J. Roy. Meteor. Soc.*, **138**, 841-861.

Pastorello, G. Z., D. Papale, H. Chu, C. Trotta, D. A. Agarwal, E. Canfora, D. D. Baldocchi, and M. S. Torn, 2017: A new data set monitors land-air exchanges. *EOS Earth & Space Science News*, **98**(8), 28-32.

Peters-Lidard, C. D., and co-authors, 2007: High performance earth system modeling with NASA/GSFC's Land Information System. *Innov. Syst. Software Eng.*, **3**, doi: 10.1007/s11334-007-0028-x.

Purdy, A. J., J. B. Fisher, M. L. Goulden, and J. S. Famiglietti, 2016. Ground heat flux: An analytical review of 6 models evaluated at 88 sites and globally. *J. Geophys. Res.,* **121**, 3045-3059.

Quiring, S. M., T. W. Ford, J. K. Wang, A. Khong, E. Harris, T. Lindgren, D. W. Goldberg, and Z. Li, 2016: North American Soil Moisture Database: Development and applications. *Bull. Amer. Meteor. Soc.*, **97**, 1441-1460.

Reichle, R. H., and Q. Liu, 2014. Observation-Corrected Precipitation Estimates in GEOS-5. NASA/TM–2014-104606, Vol. 35. http://gmao.gsfc.nasa.gov/pubs/docs/Reichle734.pdf.

Reichle, R. H., C. S. Draper, Q. Liu, M. Girotto, S. P. Mahanama, R. D. Koster, and G. De Lannoy, 2017a. Assessment of MERRA-2 land surface hydrology estimates. *J. Climate*, **30**, 2937-2960, doi: 10.1175/JCLI-D-16-0720.1.

Reichle, R., Q. Liu, R. Koster, C. Draper, S. Mahanama, and G. Partyka, 2017b. Land surface precipitation in MERRA-2. *J. Climate*, **30**, 1643-1664, doi: 10.1175/JCLI-D-16-0570.1.

Reichstein, M., and co-authors, 2005: On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. *Glob. Change Biol.*, **11**, 1424-1439, doi: 10.1111/j.1365-2486.2005.001002.x.

Rienecker, M. M., and co-authors, 2011: MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications. *J. Climate*, **24**, 3624-3648, doi:10.1175/JCLI-D-11-00015.1.

Robock, A., K. Ya. Vinnikov, C. A. Schlosser, N. A. Speranskaya and Y. Xue, 1995: Use of midlatitude soil moisture and meteorological observations to validate soil moisture simulations with biosphere and bucket models.. *J. Climate*, **8**, 15-35.

Saha, S., and co-authors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.*, **91**, 1015–1057, doi: 10.1175/2010BAMS3001.1.

Santanello, J. A., C. D. Peters-Lidard, and S. V. Kumar, 2011: Diagnosing the sensitivity of local land-atmosphere coupling via the soil moisture-boundary layer interaction. *J. Hydrometeor.*, **12**, 766-786.

Santanello, J. A., P. A. Dirmeyer, C. R. Ferguson, K. L. Findell, A. B. Tawfik, A. Berg, M. B. Ek, P. Gentine, B. Guillod, C. van Heerwaarden, J. Roundy, and V. Wulfmeyer, 2017: Land-atmosphere interactions: The LoCo perspective. *Bull. Amer. Meteor. Soc.*, (in revision).

Sellers, P. J., F. G. Hall, G. Asrar, D. E. Strebel, and R. E. Murphy, 1992: An overview of the First International Satellite Land Surface Climatology Project (ISLSCP) Field Experiment (FIFE). *J. Geophys. Res.*, **97**, 18,345-18,372.

Sellers, P. J., and co-authors, 1995: The Boreal Ecosystem-Atmosphere Study (BOREAS): An overview and early results from the 1994 field year. *Bull. Amer. Meteor. Soc.*, **76**, 1549-1577.

Sheffield, J., G. Goteti, and E. F. Wood, 2006: Development of a 50-yr high-resolution global dataset of meteorological forcings for land surface modeling. *J. Climate*, **19**, 3088-3111.

Shukla, R. P., B. Huang, L. Marx, J. L. Kinter and C.-S. Shin, 2017: Predictability and prediction of Indian summer monsoon by CFSv2: implication of the initial shock effect. *Climate Dy*n. (published online), doi: 10.1007/s00382-017-3594-0.

Sippel, S., J. Zscheischler, M. D. Mahecha, R. Orth, M. Reichstein, M. Vogel, and S. I. Seneviratne, 2017: Refining multi-model projections of temperature extremes by evaluation against land–atmosphere coupling diagnostics. *Earth Sys. Dyn.*, **8**, 387-403.

Slater, A. G., 2016: Surface solar radiation in North America: A comparison of observations, reanalyses, satellite, and derived products. *J. Hydrometeor.*, **17**, 401-420.

Viovy, N., 2013. CRUNCEP data set for 1901–2010, [Available at https://www.earthsystemgrid.org/dataset/ucar.cgd.ccsm4.CRUNCEP.v4.html].

Vuichard, N., and D. Papale, 2015: Filling the gaps in meteorological continuous data measured at FLUXNET sites with ERA-Interim reanalysis. *Earth Sys. Sci. Data*, **7**, 157-171, doi: 10.5194/essd-7-157-2015.

Williams, M., and co-authors, 2009: Improving land surface models with FLUXNET data. *Biogeosci.* **6**, 1341-1359.

Xia, Y., and co-authors, 2012: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products, *J. Geophys. Res*., **117**, D03109, doi:10.1029/2011JD016048.

Xie, P., and P. A. Arkin, 1997: Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bull. Amer. Meteor. Soc*., **78**, 2539-2558.

Xie, P., M. Chen, A. Yatagai, T. Hayasaka, Y. Fukushima, and S. Yang, 2007: "A gauge-based analysis of daily precipitation over East Asia." *J. Hydrometeor*., **8**, 607–626.

Zaitchik, B., F., J. A. Santanello, S. V. Kumar, and C. D. Peters-Lidard, 2013: Representation of soil moisture feedbacks during drought in NASA Unified WRF (NU-WRF). *J. Hydrometeor*., **14**, 360-367.

Table 1. Specifications for the four land and atmosphere model systems, including time span of data and spatial resolution. Two-letter abbreviations are used in subsequent figures and tables; generally for the first letter: N=NCEP, M=NASA (MERRA system), C=NCAR (Community models), E=ECMWF; for the second letter: L=LSM run “offline”, C=LSM coupled to GCM, R=reanalysis (except that two MERRA reanalyses are included, so they are labeled 1 and 2).

System	Offline LSM	Free-Running	Reanalysis
NOAA/ NCEP	**NL:** Noah2.7.1 [1982-2010] 1°x1° with forcing from Sheffield et al. (2006)	**NC:** CFSv2 [48 years] ~0.94°x0.94° fully coupled Shukla et al. (2017)	**NR:** CFSR [1979-2009] 0.31°x0.37° Saha et al. (2010)
NASA/ GMAO	**ML:** Catchment with boundary conditions from Mahanama et al (2015) plus physics changes [1980-2015] 0.625°x0.5° with MERRA-2 forcing and corrected precipitation Reichle et al. (2017b), GMAO (2015a,b)	**MC:** GEOS5 Heracles-5 4 p3-M3; LSM as in **ML** [2000-2015] 0.5°x0.5° with observed SST	**M2:** MERRA-2 [1980-2015] 0.625°x0.5° Gelaro et al. (2017); **M1:** MERRA [1980-2015] 0.667°x0.5° Rienecker et al (2011)
NCAR	**CL:** CLM4.5 [1991-2010] 1.25°x0.9° with CRUNCEP (Viovy 2013) forcing Lawrence et al. (2011)	**CC:** CESM 1.2.2 (CAM4 + CLM4.5) [1991-2014] 1.25°x0.9° with climatological SST	--none--
ECMWF	**EL:** HTESSEL 43R1 [1979-2015] TCo639 16km Balsamo et al. (2015)	**EC:** IFS in Athena Project [1961-2007] T1279 interpolated to N80 1.125°x1.125° with observed SST Kinter et al. (2013)	**ER:** ERA-Interim [1979-2015] 0.75°x0.75° Dee et al. (2011)

Table 2: The average value of the two terms used to calculate the terrestrial and atmospheric coupling indices using data from FLUXNET2015, each model, and averages from various groupings of the models.

	Terrestrial		Atmospheric	
	r(SM,LHF)	σ(LHF)	r(SHF,LCL)	σ(LCL)
FLUXNET2015	**0.07**	**21.2** Wm^{-2}	**0.35**	**432** m
NL	0.31	18.2	-0.22	221
NC	0.21	21.5	0.13	412
NR	0.22	23.1	0.21	396
ML	0.14	15.9	0.08	366
MC	0.13	14.0	0.02	291
M2	0.11	21.4	0.12	287
M1	0.21	22.1	0.18	340
CL	0.28	19.1	0.24	191
CC	0.18	24.1	0.15	357
EL	0.11	21.6	0.09	371
EC	0.19	17.7	0.08	350
ER	0.05	18.8	0.13	291
All	**0.18**	**19.8**	**0.10**	**323**
LSMs	0.21	18.7	0.05	287
Coupled	0.18	19.3	0.10	352
Reanalyses	0.15	21.4	0.16	328

Figure Captions:

Figure 1: Location of the FLUXNET2015 Tier-1 sites used in this study. Triangles indicate no upward shortwave radiation measurements available to estimate surface albedo, pluses mean no Bowen ratio corrected surface heat fluxes provided, exes indicate neither albedo nor corrected heat fluxes are available, circles have both. Color of the symbol indicates the length of data series available.

Figure 2: Scatter of annual total precipitation measurements at FLUXNET2015 sites (abscissa) to estimates (ordinate) from gridded observationally-based precipitation analyses (top two rows) or reanalyses constrained by data assimilation (bottom row) using the value from the grid box containing the FLUXNET2015 site location (unless data are missing or indicated to be an all-ocean grid box). Dash-dotted diagonal grey line indicates X=Y. Colors indicate years of available data from each FLUXNET2015 site, whiskers span range of annual totals from FLUXNET2015 (horizontal) or gridded estimates (vertical) for years where data sets overlap. Purple line is the best-fit linear regression of Y on X. Statistics are explained in the text.

Figure 3: Taylor diagram of annual mean surface radiation terms for the 12 indicated models verified against FLUXNET2015 sites for downward solar radiation (black), downward longwave radiation (red), upward shortwave radiation (blue) and net radiation (green). Dot colors indicate mean bias and size shows percentage of stations where the range of the annual totals from the model overlaps the span from FLUXNET2015 sites (also presented in tabular form in the upper right).

Figure 4: As in Fig. 3 for surface albedo; annual mean (black) and boreal summer (JJA) mean (red).

Figure 5: Propagation of errors estimated from their rank correlations among precipitation (P), height of the lifting condensation level (LCL), evaporative fraction (EF), sensible and latent heat flux (SH & LH), surface albedo (α) and net radiation (R_{Net}) across FLUXNET2015 stations. Ratios show the number of models out of 11 (correlations involving α) or 12 (other variables) with p-values below 0.10; p-value shown is based on the average of correlations across all models. Widths of arrows follow significance of correlations and no arrows are drawn where p-values are large. The wide double arrows between EF and heat fluxes denote p-values $< 10^{-12}$.

Figure 6: As in Fig. 3 for the magnitude of the annual cycle (first harmonic calculated from monthly means) of sensible heat flux (orange), latent heat flux (cyan) and net radiation at the surface (green).

Figure 7: As in Fig. 6 for phase of the annual cycle of sensible heat flux (orange) and latent heat flux (cyan) and net radiation at the surface (green).

Figure 8: As in Fig. 6 for the magnitude (brown) and phase (purple) of the annual cycle of surface albedo.

Figure 9: Distribution of coupling indices for the terrestrial (x-axis) and atmospheric (y-axis) legs for the warmest consecutive 3 months of the annual cycle for FLUXNET2015 sites (white dots; identical in each panel) and for each model as indicated. Colors of dots indicate in which quadrant that FLUXNET2015 site lies: red = both indices positive; green = terrestrial positive, atmospheric negative; blue = atmospheric positive, terrestrial negative; grey = both negative. The white circle indicates the centroid of all FLUXNET2015 stations that are in that quadrant, connected by a colored dotted line to a colored circle that is the centroid of the same stations' corresponding grid boxes as simulated by the model. Numbers in the corners of each quadrant show the number of

points in that quadrant according to the model and FLUXNET2015 data, separated by a colon, and the percentage of the FLUXNET2015 sites within that quadrant that the model placed in the same quadrant. The percentage in red at the upper right of each panel is the overall percentage of sites where model and FLUXNET2015 agree on the quadrant.

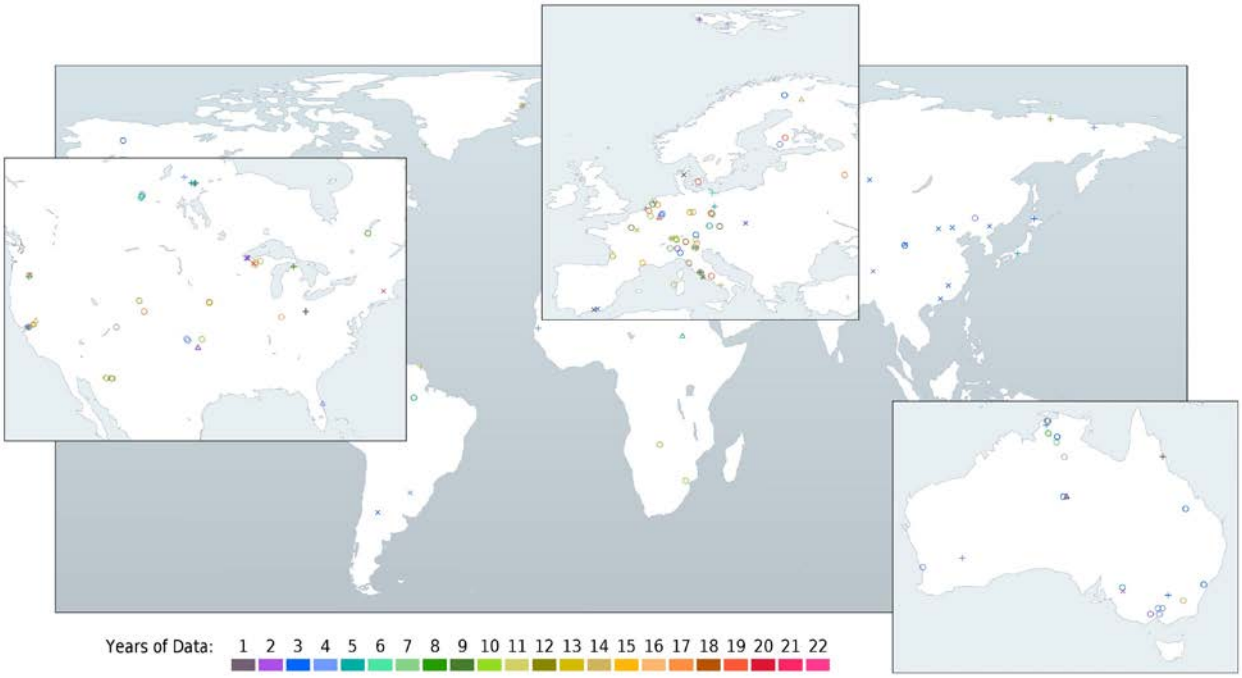
Years of Data: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

Figure 1: Location of the FLUXNET2015 Tier-1 sites used in this study. Triangles indicate no upward shortwave radiation measurements available to estimate surface albedo, pluses mean no Bowen ratio corrected surface heat fluxes provided, exes indicate neither albedo nor corrected heat fluxes are available, circles have both. Color of the symbol indicates the length of data series available.

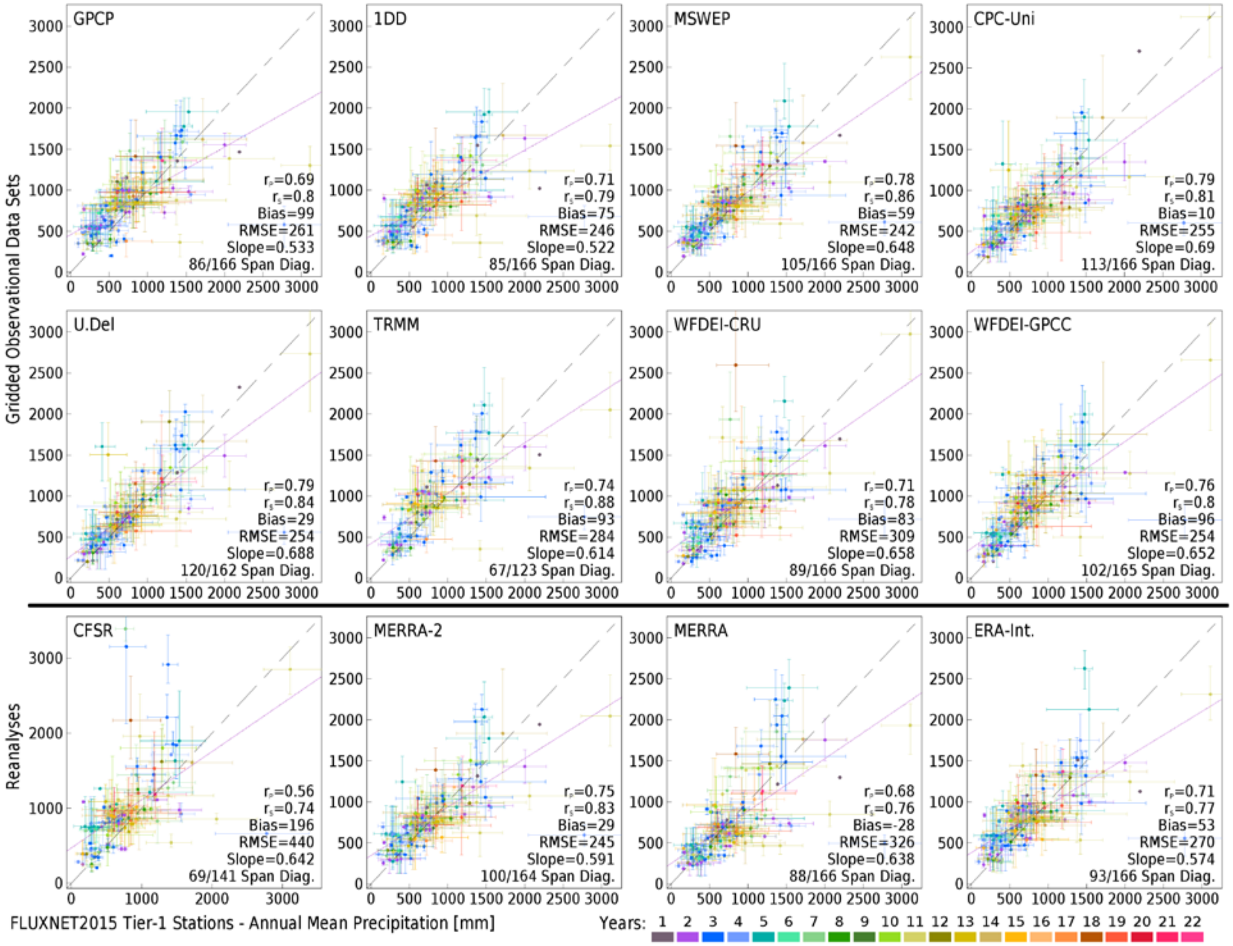

Figure 2: Scatter of annual total precipitation measurements at FLUXNET2015 sites (abscissa) to estimates (ordinate) from gridded observationally-based precipitation analyses (top two rows) or reanalyses constrained by data assimilation (bottom row) using the value from the grid box containing the FLUXNET2015 site location (unless data are missing or indicated to be an all-ocean grid box). Dash-dotted diagonal grey line indicates X=Y. Colors indicate years of available data from each FLUXNET2015 site, whiskers span range of annual totals from FLUXNET2015 (horizontal) or gridded estimates (vertical) for years where data sets overlap. Purple line is the best-fit linear regression of Y on X. Statistics are explained in the text.

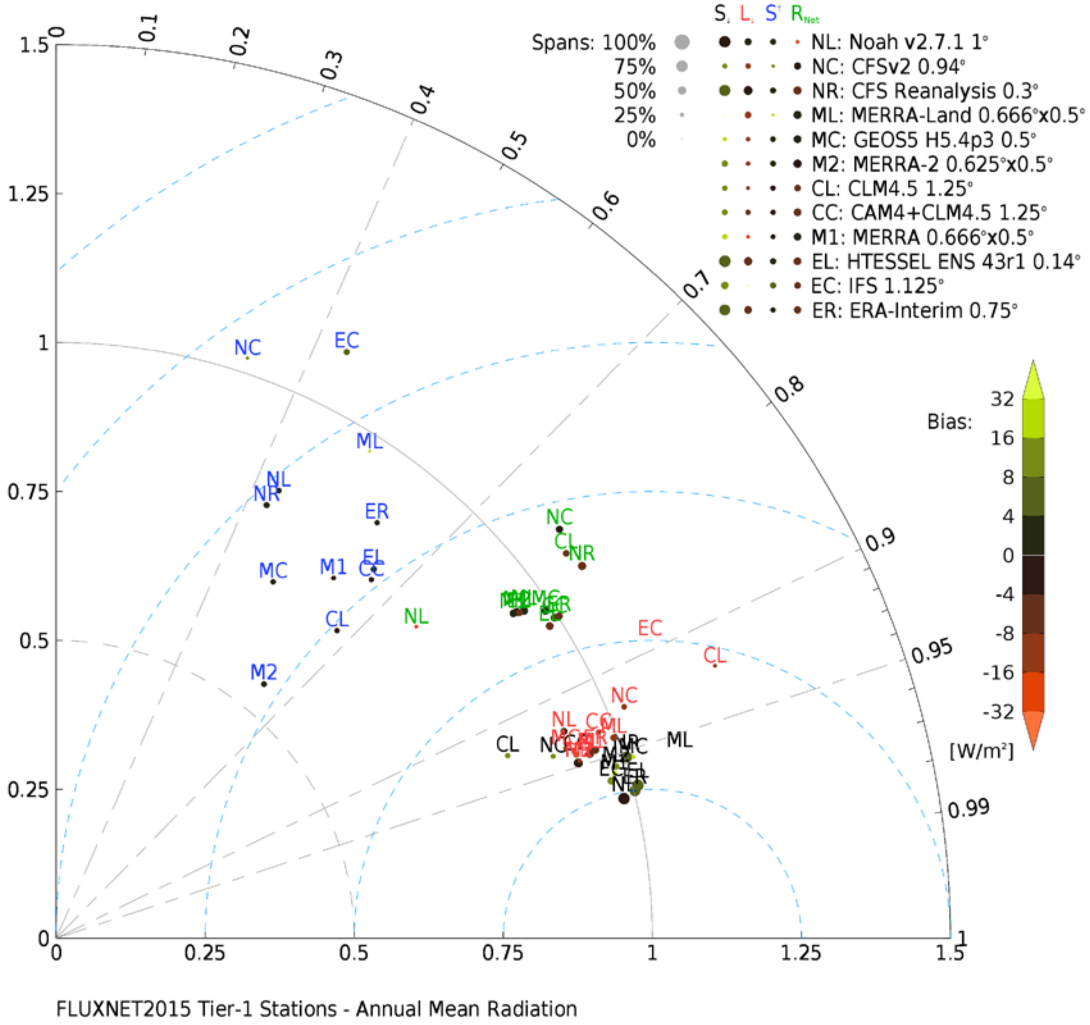

Figure 3: Taylor diagram of annual mean surface radiation terms for the 12 indicated models verified against FLUXNET2015 sites for downward solar radiation (black), downward longwave radiation (red), upward shortwave radiation (blue) and net radiation (green). Dot colors indicate mean bias and size shows percentage of stations where the range of the annual totals from the model overlaps the span from FLUXNET2015 sites (also presented in tabular form in the upper right).

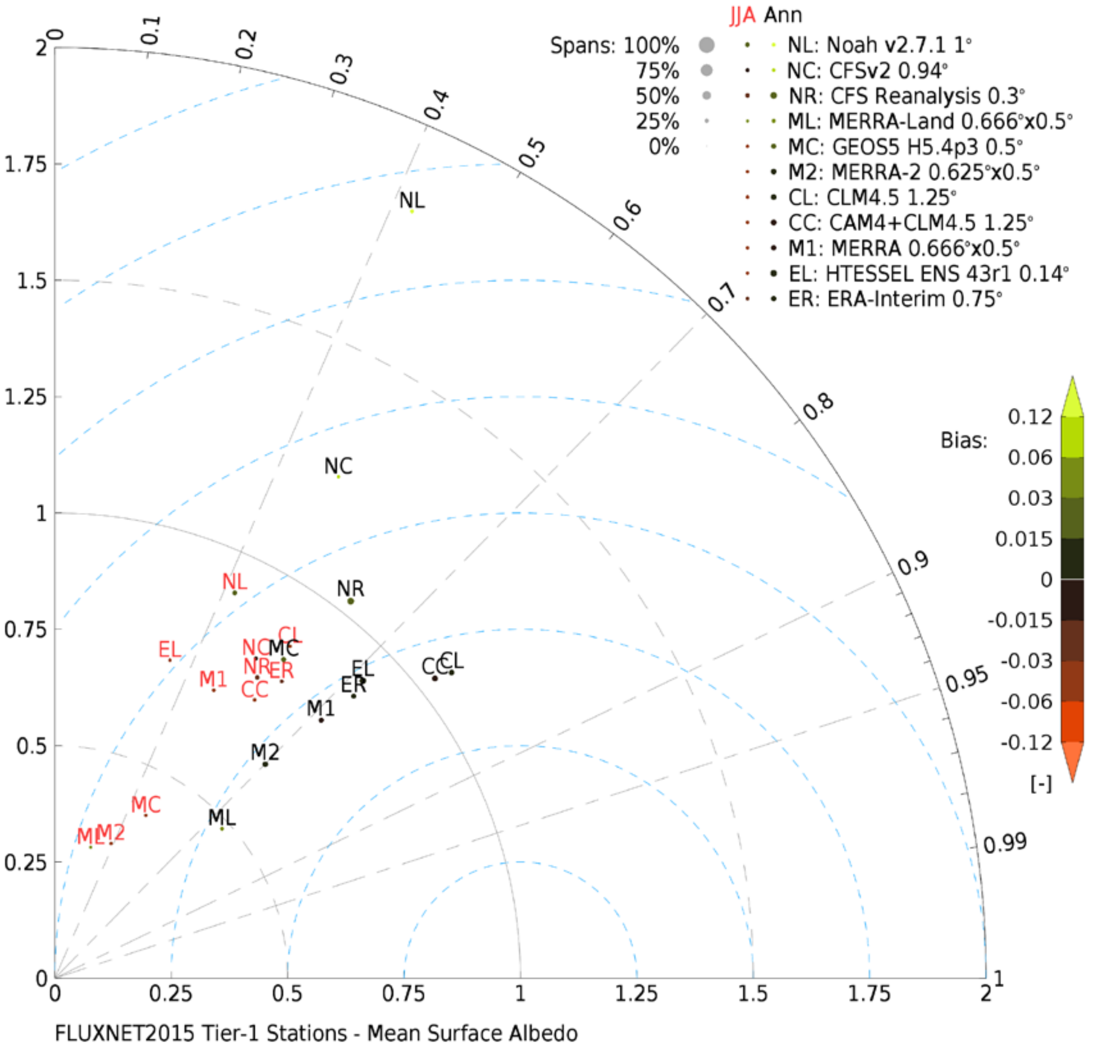

Figure 4: As in Fig. 3 for surface albedo; annual mean (black) and boreal summer (JJA) mean (red).

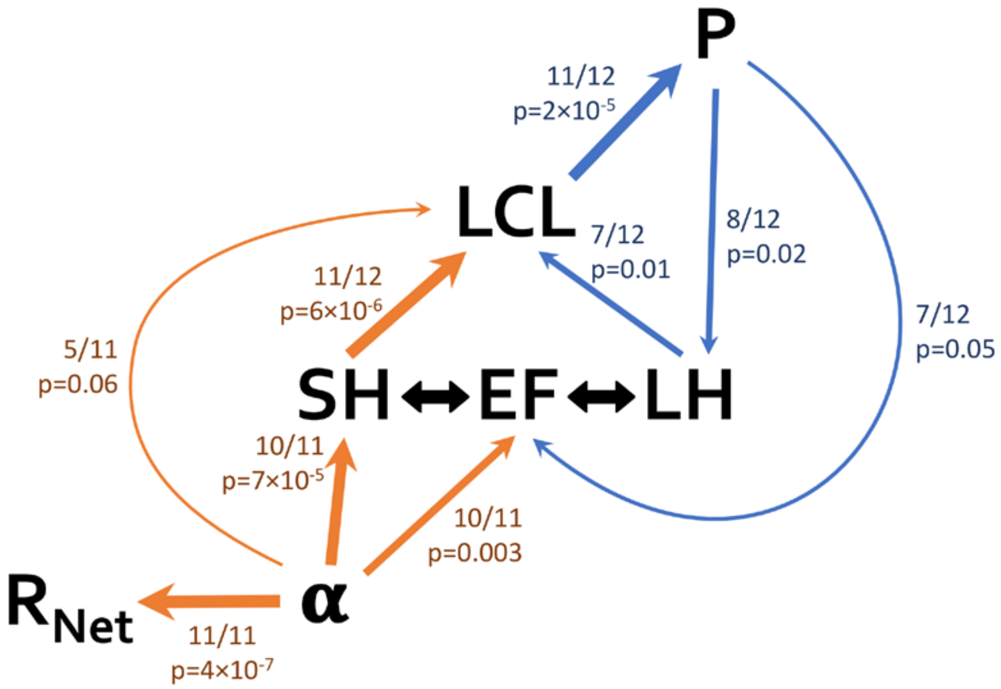

Figure 5: Propagation of errors estimated from their rank correlations among precipitation (P), height of the lifting condensation level (LCL), evaporative fraction (EF), sensible and latent heat flux (SH & LH), surface albedo (α) and net radiation (R_{Net}) across FLUXNET2015 stations. Ratios show the number of models out of 11 (correlations involving α) or 12 (other variables) with p-values below 0.10; p-value shown is based on the average of correlations across all models. Widths of arrows follow significance of correlations and no arrows are drawn where p-values are large. The wide double arrows between EF and heat fluxes denote p-values $< 10^{-12}$.

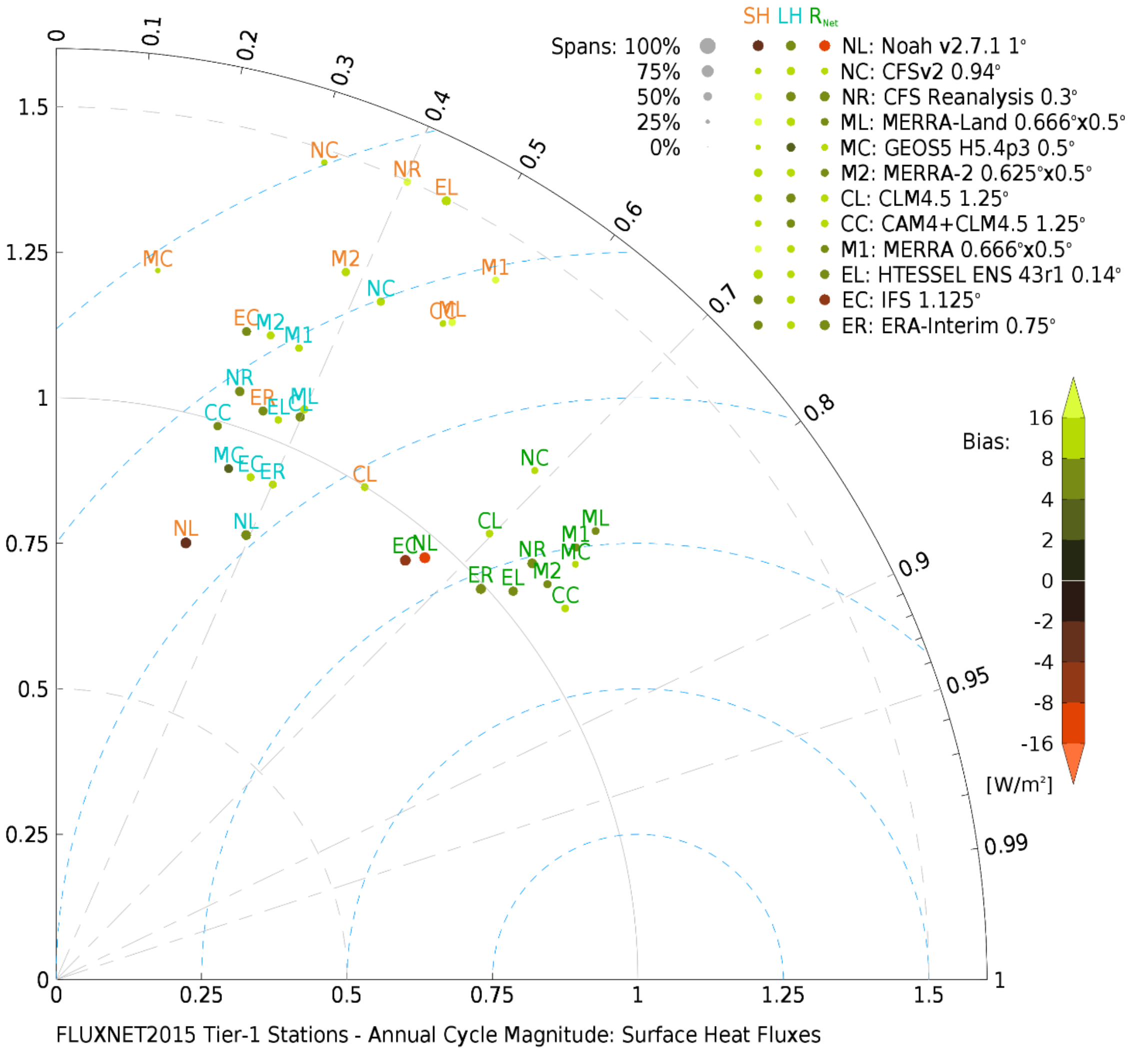

Figure 6: As in Fig. 3 for the magnitude of the annual cycle (first harmonic calculated from monthly means) of sensible heat flux (orange), latent heat flux (cyan) and net radiation at the surface (green).

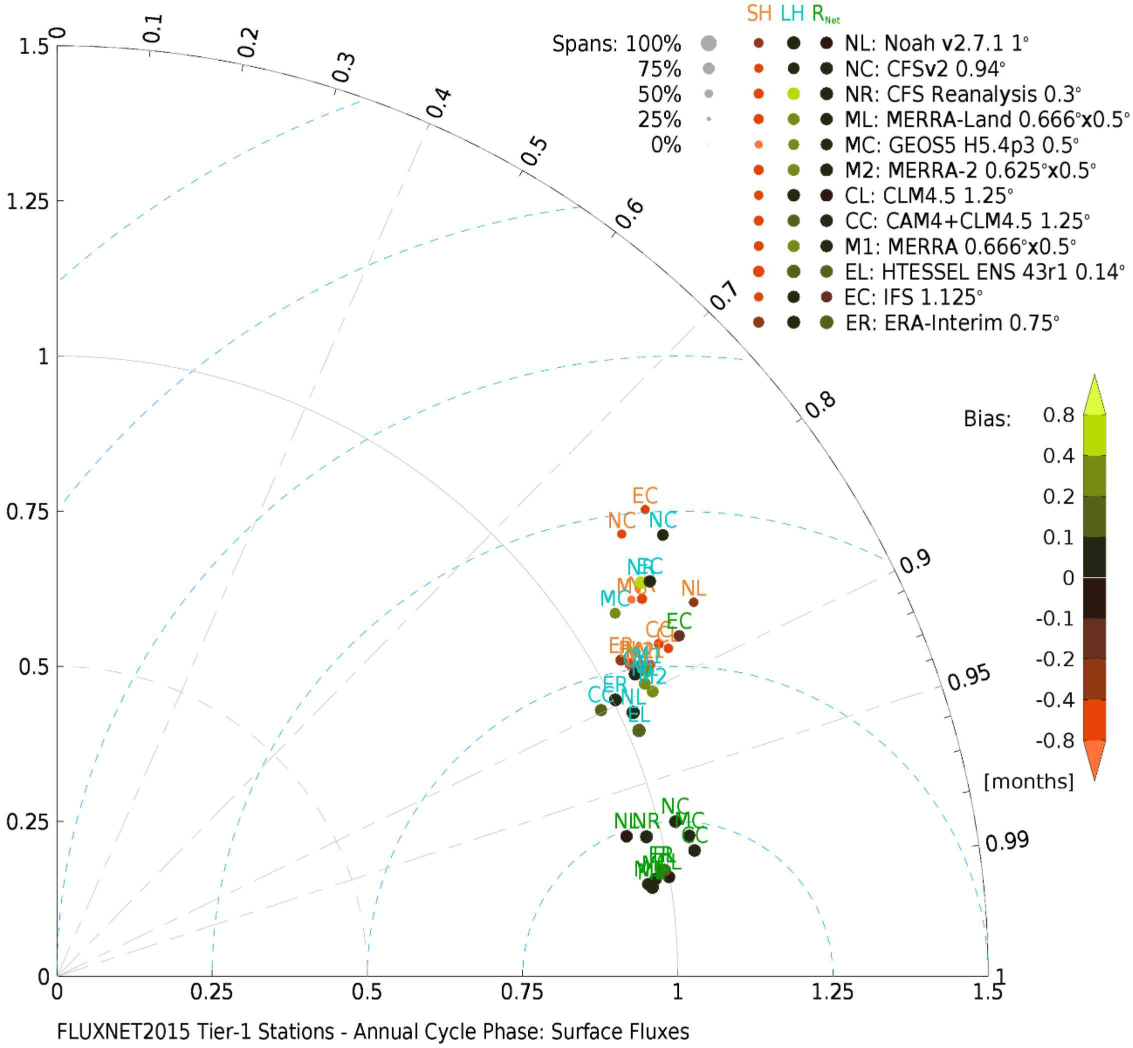

Figure 7: As in Fig. 6 for phase of the annual cycle of sensible heat flux (orange) latent heat flux (cyan), and net radiation at the surface (green).

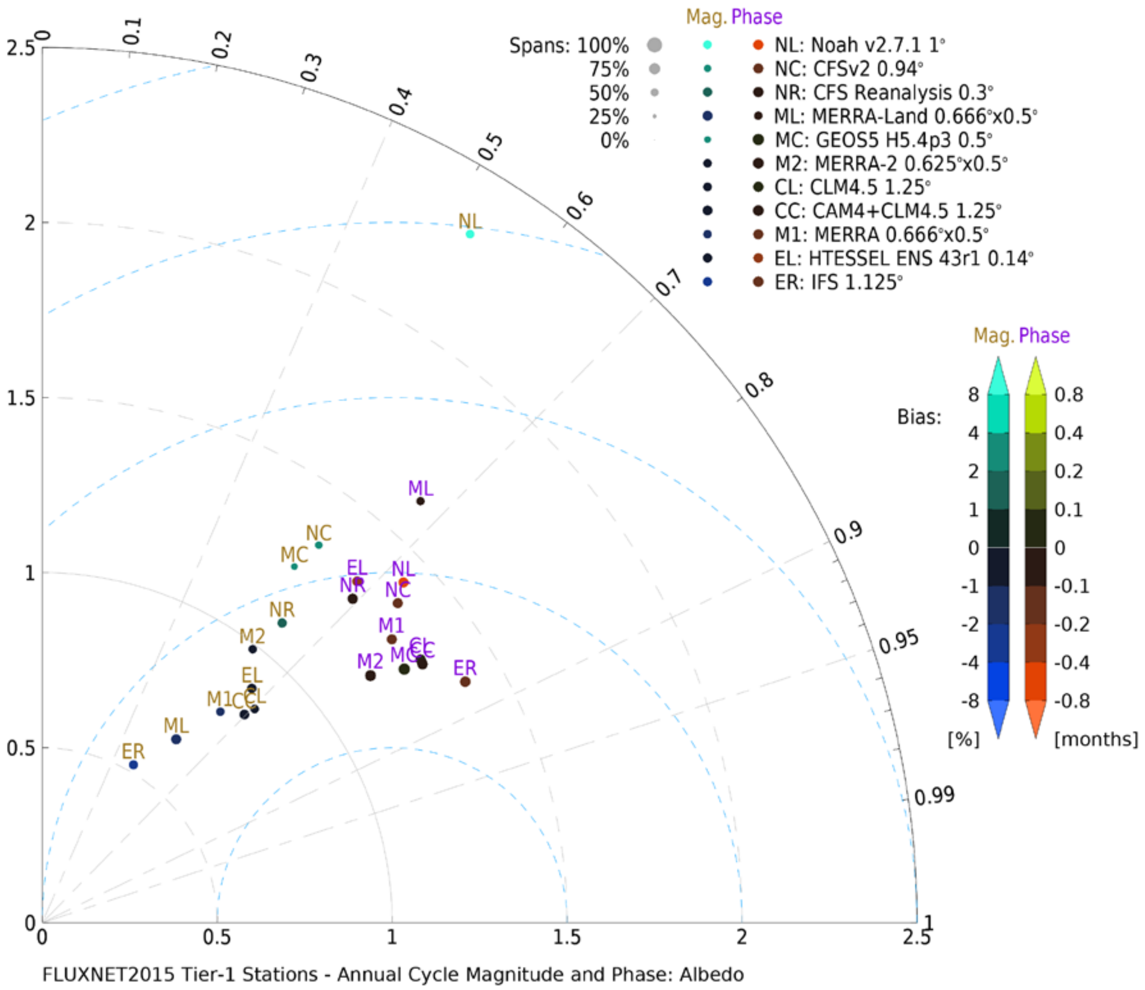

Figure 8: As in Fig. 6 for the magnitude (brown) and phase (purple) of the annual cycle of surface albedo.

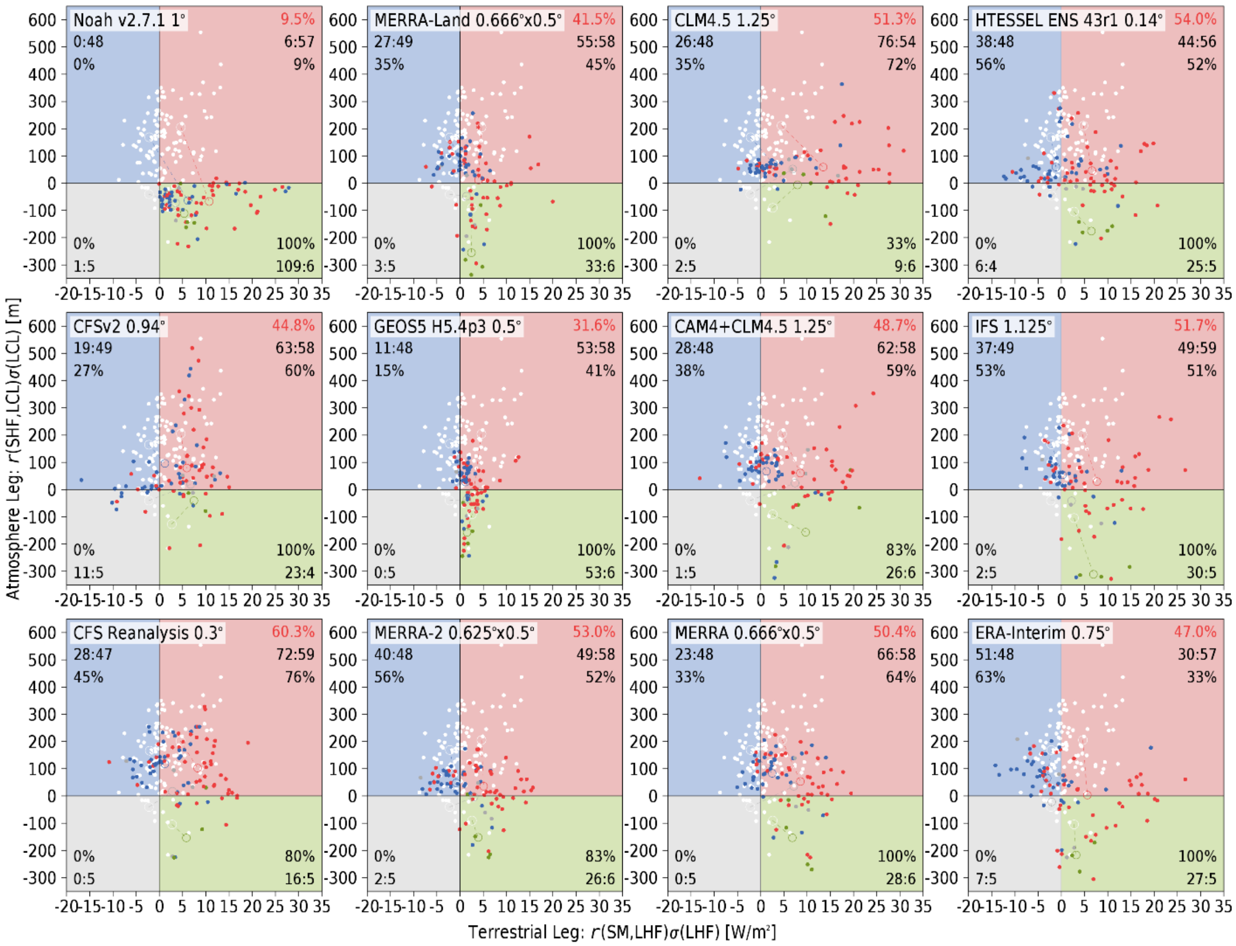

Figure 9: Distribution of coupling indices for the terrestrial (x-axis) and atmospheric (y-axis) legs for the warmest consecutive 3 months of the annual cycle for FLUXNET2015 sites (white dots; identical in each panel) and for each model as indicated. Colors of dots indicate in which quadrant that FLUXNET2015 site lies: red = both indices positive; green = terrestrial positive, atmospheric negative; blue = atmospheric positive, terrestrial negative; grey = both negative. The white circle indicates the centroid of all FLUXNET2015 stations that are in that quadrant, connected by a colored dotted line to a colored circle that is the centroid of the same stations' corresponding grid boxes as simulated by the model. Numbers in the corners of each quadrant show the number of points in that quadrant according to the model and FLUXNET2015 data, separated by a colon, and the percentage of the FLUXNET2015 sites within that quadrant that the model placed in the same quadrant. The percentage in red at the upper right of each panel is the overall percentage of sites where model and FLUXNET2015 agree on the quadrant.